THE FOLLOWING PAGES WERE NOT

RECEIVED WITH ORIGINAL DOCUMENT AND

THEREFORE HAVE NOT BEEN REPRODUCED:

P 16

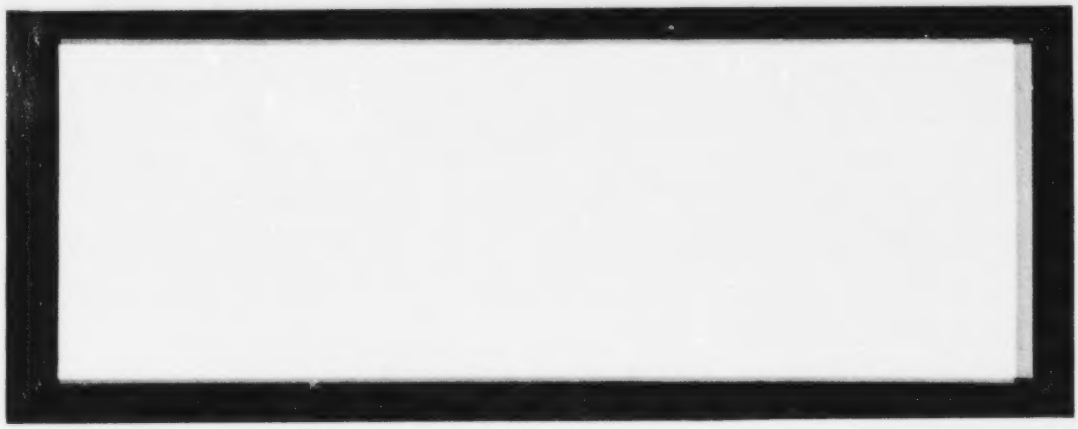
PUISQUE LES PAGES SUIVANTES N'ONT PAS ÉTÉ

REÇUES AVEC LE DOCUMENT ORIGINAL

ELLES N'ONT PAS ÉTÉ REPRODUITES:

P. 16.

Statistics Canada | Statistique Canada



Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

Z-157 E
c.2

A REVIEW OF THE EDIT AND IMPUTATION SYSTEM FOR THE 100% TAX FILE

Mark Armstrong and Edouard Auger

Working Paper No. CHSMΩ 85-030-E

TABLE OF CONTENTS

		PAGE
1.	INTRODUCTION	1
2.	OBJECTIVE OF E/I	2
3.	THE E/I SYSTEM FOR THE 100% TAX FILE	3
4.	THE EFFECT OF E/I	6
5.	QUALITY REPORTING ON THE E/I	7
6.	CONCLUSIONS	8
APPENDIX A	EFFECTS OF IMPUTATION	9
	TABLE A.1 Family Allowance Income by Marital Status.	9
	TABLE A.2 Old Age Security Benefit Income by Gender.	10
	TABLE A.3 Self-employment Income by Marital Status..	10
	TABLE A.4 Employment Income by Gender	11
	TABLE A.5 Total Income	11
APPENDIX B	EFFICIENCY MEASURES FOR THREE YEARS	13
APPENDIX C	QUALITY REPORTING FOR 1981, 1982 AND 1983 TAX YEARS..	15
	TABLE C.1 Date of Birth	16
	TABLE C.2 Marital Status	16
	TABLE C.3 Gender	16
	TABLE C.4 Marital Status by Gender	17
	TABLE C.5 Total Income by Gender	18
	TABLE C.6 Gross and Net Incomes	20

1. INTRODUCTION

Since 1979, the Administrative Data Development Division (ADDD) of Statistics Canada has received from Revenue Canada Taxation (RCT) a computer tape of the 100% tax file. [1] This incoming tape contains geographic, demographic and economic information as well as an identification variable for every Canadian taxfiler. The information is confidential and access is permitted to only a couple of persons within ADDD.

The data received from RCT is not clean in a sense that errors and ambiguities exist. A record will sometimes lack a value for a certain variable or give values which appear to be invalid. Inconsistencies in the relationships between variables do appear and mathematical mistakes occur.

The purpose of this report is to examine the edit and imputation (E/I) system developed at STC from a more-or-less heuristic view. This report will focus on the objectives of an E/I system, the concepts involved, its impact and significance. Some methodology will be discussed but the purer technical aspects such as the computer system operations will not be covered.

---

[1] ADDD does however hold tax data since the 1976 tax year.

2. *OBJECTIVES OF E/I*

The purpose of E/I is to eliminate, or at least recognize, the types of errors and ambiguities previously stated and to monitor the distributions of some variables of interest. The end result should be a computer taxfile that is cleaner than the original incoming file from RCT. Data analysis will show more consistency in totals, and bias due to having to eliminate records with non-response will be reduced. The E/I system was not created to handle records of total non-response. Invalid data will be resolved and become usable data in the calculation of statistics derived from the taxfile.

Because of the complexity of the Canadian taxation system and the number of persons filing returns, the E/I could consume a large amount of computer and human resources. One of the key considerations in developing and maintaining an E/I system was that it was cost-efficient in both resource areas.

3. *THE E/I SYSTEM FOR THE 100% TAX FILE*

The E/I system handles roughly 15,000,000 tax records twice each year. The preliminary and final files are processed annually wherein each individual record is independently processed. [2] Variables which are used in the E/I include marital status, gender, day/month and year of birth, and income split into types like total, gross net, commission, self-employment, employment, family allowance, and old age security benefits.

In developing the E/I system, a test file was used rather than a complete tax file. The test data was a 1979 file on New Brunswick personal identification/personal income variables similar to those on the tax file. The proportion of errors found in this study file was quite small and the edits and imputations have been based on this. It was decided that no elaborate E/I system needed to be devised to adequately perform on the tax data. As a result of testing editing and imputation algorithms and options on the New Brunswick file, a simple and straight forward and relatively low cost E/I system was built to handle tax records.

The E/I system usually requires some program adjustments for each new tax year. Changes to the record layout and to tax legislation would require that some work be done before the tax file is put through the system. Adjustments to the edit and imputation parameters are done automatically during the previous years tax file processing.

---

[2] The difference between the preliminary and final file is that the final file contains late returns. As well, the information for an individual return could change between the preliminary and final file. The number of records that are added onto the final file is not constant over the years. For 1982, the percent increase in the number of records between preliminary and final file was 1.1%.

Editing and Imputation

In planning and devising edit and imputation schemes for the E/I, several fundamental assumptions were made. For example, the quality of the income information was probably better than demographic or other information because of the processing done at RCT. Various other assumptions were made when there was a conflict between demographic variable information. For instance, if the marital status indicated was "single" but there was a spouse's SIN, it was assumed that the taxfiler was previously married at the tax date.

The edits are fairly straight forward. Rounding dollars and cents to just dollars and adjusting for "impossible" negative values to positive values are two such edits. Decision logic tables were an integral part of the E/I system development. Tables were constructed for the following variables, day and month of birth, year of birth and marital status. Economic data was used to a large extent as auxiliary information in the decision making process.

Before the main E/I program is run, there is a pre-edit stage. At this point, limits are established for 'extreme value detection' and various tables are produced for geographic and demographic variables. For economic variables, Tukey's method of estimating inner and outer limits are employed. This method is quite robust and uses quartiles rather than means and standard deviations to determine boundaries which could establish extreme values.

With such a method of outlier detection, observations of extreme likelihood can be determined and extracted from the tax file for subject-matter investigation. Corrections to outliers are done manually by subject-matter persons.

As part of the main E/I program, imputation reconciles total income and its many components as well as confirming the calculation of employment expenses.

Imputation is really the heart of any E/I system. In developing an imputation strategy it was felt that the distributions of different variables be preserved against the introduction of new and

different errors.

Probabilistic imputation selects values at random from empirical distributions obtained from the valid and consistent information. These distributions use information from auxiliary variables, when applicable, and are suitable for variables needing imputation. Various combinations of these auxiliary variables are used in order to determine the most satisfactory predictor variables to be used in imputation.

4. THE EFFECT OF E/I

With the quantity of missing, invalid, and inconsistent information in the tax file being quite small, the imputations do not have a large effect on the final estimates. Imputations occur only in about 1% of the population.

Many different variables were studied to see what the effect of E/I was. In the study of income variables, cross classified by demographic variables, the median values with imputed and non-imputed were the same. The mean values had some variance in repeated imputation trials but the differences between imputed and non-imputed values were neither statistically or subject-matter significant. Appendix A illustrates some comparisons of imputed and non-imputed data.

5. *QUALITY REPORTING ON THE E/I*

The E/I program produces a monitoring or quality report. This report looks at the demographic and economic variables in tabular form. Tables show the data summarized according to before and after E/I. Efficiency measures for demographic variables were used to evaluate the imputations of missing or invalid information. Based on these measures, the imputation procedures proved to be quite satisfactory. Appendix B describes these measures.

Monitoring reports have been looked at for 1981, 1982 and 1983 to see whether there are significant distributional changes occuring over time. Although some levels of a couple of variables are showing changes over the yearly periods, the changes are quite small, usually in the order of 0.1%. Some tables have been included in Appendix C showing the values and changes of some variables on the quality reports for the past three years.

6. CONCLUSIONS

The E/I system for the tax file is efficient in resources while meeting the needs of both data analysts and subject-matter persons in ADDD. It is a relatively simple system but it does require a bit of manual work each year. The preliminary tax information is received about 6 months after the end of April and a final version follows from RCT about one year after that. The E/I system can be run almost as soon as the tapes arrive.

The E/I is not a static piece of software and plans are in the making for major revisions based upon current work within ADDD. These revisions will add to the system and be able to examine the family relationships from a tax perspective.

APPENDIX A: EFFECTS OF IMPUTATION

Imputation has some effect on the final estimates compared to the estimates without imputation. The induced variability caused by E/I is of a great concern, especially since distributional changes might occur. With the sample sizes involved, however, the impact of imputation is quite undiscernable. The tables below review the impact on different types of income estimates after imputing for missing and invalid information.

The source of the data in the following tables was the 1979 personal identification/income file for New Brunswick. Three imputation trials were made for each variable studied. The mean of these three trials is also given.

The units of "sum", "median" and "mean" are dollar values based on the number of taxfilers (N).

In general, there are some very small fluctuations in the mean and median values but they are negligible. The estimates for the imputed data are quite stable and very close to the non-imputed values. Thus, the imputation methods are very good for the variables looked at.

TABLE A.1 FAMILY ALLOWANCE INCOME BY MARITAL STATUS

		MARITAL STATUS		
		MARRIED	SINGLE	OTHER
NON-IMPUTED	N	96,879	4,478	12,237
	SUM	46,517,795	1,229,276	5,273,569
	MEAN	480.164	274.513	430.953
	MEDIAN	480	240	400
IMPUTED 1	N	96,892	4,294	12,443
	SUM	46,524,696	1,148,967	5,358,712
	MEAN	480.171	267.575	430.661
	MEDIAN	480	240	400
IMPUTED 2	N	96,894	4,295	12,440
	SUM	46,524,756	1,149,127	5,358,492
	MEAN	480.161	267.55	430.747
	MEDIAN	480	240	400
IMPUTED 3	N	96,894	4,297	12,438
	SUM	46,525,336	1,150,087	5,356,952
	MEAN	480.167	267.649	430.692
	MEDIAN	480	240	400
MEAN OF TRIALS	N	96,893.3	4,295.3	12,440.3
	SUM	46,524,929.3	1,149,393.6	5,358,052
	MEAN	480.166	267.591	430.7
	MEDIAN	480	240	400

TABLE A.2 OLD AGE SECURITY BENEFIT INCOME BY GENDER

		SEX	
		MALE	FEMALE
NON-IMPUTED	N	16,253	11,325
	SUM	31,830,360	22,451,811
	MEAN	1,958.43	1,982.5
	MEDIAN	2,074	2,074
IMPUTED 1	N	16,330	11,577
	SUM	31,978,334	22,955,410
	MEAN	1,958.26	1,982.85
	MEDIAN	2,074	2,074
IMPUTED 2	N	16,328	11,579
	SUM	31,973,675	22,960,006
	MEAN	1,958.21	1,982.91
	MEDIAN	2,074	2,074
IMPUTED 3	N	16,326	11,581
	SUM	31,975,800	22,957,944
	MEAN	1,958.58	1,982.38
	MEDIAN	2,074	2,074
MEAN OF TRIALS	N	16,328	11,579
	SUM	31,975,936.3	22,957,807.7
	MEAN	1,958.35	1,982.68
	MEDIAN	2,074	2,074

TABLE A.3 SELF-EMPLOYMENT INCOME BY MARITAL STATUS

		MARITAL STATUS		
		MARRIED	SINGLE	OTHER
IMPUTED 1	N	25,460	3,367	1,967
	SUM	142,986,736	12,383,305	7,573,105
	MEAN	5,616.13	3,677.85	3,850.08
	MEDIAN	2,223	1,742	1,743
IMPUTED 2	N	25,460	3,369	1,965
	SUM	142,986,736	12,399,803	7,556,607
	MEAN	5,616.13	3,680.56	3,845.6
	MEDIAN	2,223	1,744	1,740
IMPUTED 3	N	25,461	3,369	1,964
	SUM	142,991,032	12,399,048	7,556,780
	MEAN	5,616.08	3,679.44	3,847.28
	MEDIAN	2,224	1,742	1,741.5
MEAN OF TRIALS	N	25,460.3	3,368.3	1,965.1
	SUM	142,988,168	12,393,052	7,561,024
	MEAN	5,616.11	3,679.28	3,847.65
	MEDIAN	2,223.7	1,742.3	1,741.1

TABLE A.4 EMPLOYMENT INCOME BY GENDER

		SEX	
		MALE	FEMALE
IMPUTED 1	N	188,798	114,990
	SUM	2,290,243,482	732,216,647
	MEAN	12,130.7	6,367.65
	MEDIAN	11,077	5,043
IMPUTED 2	N	188,814	114,974
	SUM	2,290,305,386	732,154,743
	MEAN	12,130	6,368
	MEDIAN	11,076.5	5,044
IMPUTED 3	N	188,805	114,983
	SUM	2,290,258,118	732,202,011
	MEAN	12,130.3	6,367.92
	MEDIAN	11,077	5,043
MEAN OF TRIALS	N	188,805.7	114,982.3
	SUM	2,290,268,995.3	732,191,133.7
	MEAN	12,130.3	6,367.86
	MEDIAN	11,076.83	5,043.7

TABLE A.5 TOTAL INCOME

		TOTINC
NON-IMPUTED	N	280,697
	SUM	4,439,425,357
	MEAN	15,815.7
	MEDIAN	12,773
IMPUTED 1	N	280,697
	SUM	4,439,205,123
	MEAN	15,814.9
	MEDIAN	12,772
IMPUTED 2	N	280,697
	SUM	4,439,205,436
	MEAN	15,814.9
	MEDIAN	12,772
IMPUTED 3	N	280,697
	SUM	4,439,205,084
	MEAN	15,814.9
	MEDIAN	12,772
MEAN OF TRIALS	N	280,697
	SUM	4,439,205,214.3
	MEAN	15,814.9
	MEDIAN	12,772

CONCLUSIONS

The five tables show that the estimates of different income variables (mean and median) do not change significantly after imputation is made.

APPENDIX B: EFFICIENCY MEASURES

Efficiency measures are produced in the monitoring reports to indicate how successful imputing was. Efficiency measures are calculated for missing values of the year of birth (excluding those persons with OAS benefits), marital status and gender.

The efficiency measures are calculated as follows

$$\text{Year of Birth Efficiency} = \frac{\sum_{j=1}^{6} \left(\sum_{i=1}^{3 \text{ or } 6} (Pij \times Plij) \right) nj}{\sum_{j=1}^{6} nj}$$

where i=1,2,3 or i=1,2,3,4,5,6

These correspond to the following respective age groups:

15-34, 35-54, 55+ and 15-24, 25-34, 35-54, 55-64, 65+, 14-

j = six marital status groups

Pij = the proportion in group i in category j as found in the monitoring reports

Plij = the corresponding proportion as used in the imputation table.

nj = the number of missing values in a marital status category.

$$\text{Gender Efficiency} = \frac{\sum_{j=1}^{282} \left(\sum_{i=1}^{2} (Pij \times Plij) \right) nj}{\sum_{j=1}^{282} nj}$$

where i = 1,2, (male, female)

j = 1, ..., 282 (46 total income intervals cross-classified by 6 marital status categories)

The efficiency measure for marital status is similar to the year of birth but with different imputation categories (i.e., single, married, divorced, widowed, separated, married with no spouse's SIN present.

The higher the efficiency measure, the more successful imputation has been in assigning a correct value. Table 3.1 shows the results for the three efficiency measures for 1981, 1982 and 1983. (The preliminary 1983 report was used, final reports for 1981 and 1982).

TABLE 3.1 EFFICIENCY MEASURES FOR THREE YEARS

EFFICIENCY MEASURE	1981	1982	1983
YEAR OF BIRTH			
i = 3 categories	0.59	0.58	0.62
i = 6 categories	0.43	0.34	0.35
MARITAL STATUS			
6 categories	0.87	0.68	0.61
GENDER			
6 categories	0.64	0.64	0.65

The values in Table B.1 are satisfactory since they are around 60% except for YOB when six categories of age are imputed. The measures for gender are quite stable. What these values say is that for every 100 records where no gender is indicated, E/I will accurately assign about 65 to the proper category. The other values are interpreted in a similar fashion. The rather large drop in the efficiency for marital status from the 1981 to 1982 to 1983 can be attributed to the fact that there were about 3.5 times the number of records with no marital status in 1982 as there were in 1981. In 1983, the number of records with no marital status fell again from the 1982 values.

A P P E N D I X C: QUALITY REPORTING FOR 1981,1982,1983 TAX YEARS

The annual reports since 1981 look at the values and distributions of many of the tax file variables. One main item of importance is that E/I is able to detect large changes in the distribution of say male versus female taxfilers. The quality reports, compared over time will indicate trends in the demographic and economic tax file structures. The actual values themselves are important to subject-matter persons working with tax data.

The tables in this appendix show the results of the quality reports for some variables for 1981, 1982 and 1983 along with a direction of trend, if any. The preliminary 1983 report was used since ADDD had not received the final file when this report was written. This should not be a significant consideration however.

Table C.1

DATE OF BIRTH (expressed as a percent)

	1981	1982	1983	Trend
Month of birth and date of birth valid	99.97	99.90	99.86	↓
Month of birth invalid	—	—	—	
Date of birth invalid	—	—	—	
Month of birth and date of birth invalid	0.03	0.10	0.14	↑

Table C.2

MARITAL STATUS (expressed as a percent)

	1981	1982	1983	Trend
Married	61.11	61.34	61.12	
Single	26.61	25.62	25.61	
Widow	4.90	4.85	4.98	
Divorced	3.63	3.86	4.04	↑
Separated	3.49	3.51	3.50	
Married with name of spouse blank	0.03	0.02	0.00	
Missing	0.23	0.80	0.75	

Table C.3

GENDER (before E/I, percentages)

	1981	1982	1983	Trend
Males	52.60	52.26	51.92	↓
Females	47.13	47.46	47.84	↑
Missing	0.27	0.28	0.24	

Table C.4

MARITAL STATUS X GENDER (expressed as a percent)

	MALE	FEMALE	MISSING	TRENDS?
Married				
1981	33.47	27.57	0.08	
1982	33.57	27.98	0.09	Stable for males
1983	33.25	28.06	0.07	↑ for females
Single				
1981	15.21	11.53	0.07	
1982	14.63	11.26	0.08	Stable for females
1983	14.50	11.34	0.07	↓ for males
Widow				
1981	0.91	3.89	0.11	
1982	0.88	3.93	0.10	Slow ↑ for females
1983	0.99	3.98	0.09	
Divorced				
1981	1.44	2.19	<0.01	
1982	1.58	2.33	<0.01	Gradual increase for
1983	1.62	2.47	<0.01	both genders
Separated				
1981	1.56	1.93	<0.01	
1982	1.60	1.95	<0.01	Stable for males
1983	1.62	1.99	<0.01	↑ for females
Other				
1981	0.00	0.03	<0.01	
1982	<0.01	0.01	<0.01	
1983	0.00	0.00	0.00	

Other = married but first 4 characters of spouses name is blank.

NOTE: The totals for the marital status and gender variables may not equal the totals when these variables are looked at separately. This is due to the rounding used.

Table C.5

TOTAL INCOME BY GENDER

SINGLE

INCOME	MALE	FEMALE	MISSING	TRENDS?
-1<$<44k				
1981		(not calculated)		
1982	0.26	0.09	<0.01	Percentages are
1983	0.27	0.12	<0.01	increasing
0<$<999				
1981	0.70	0.98	0.02	
1982	0.74	0.97	0.02	Percentages are
1983	0.82	1.10	0.02	increasing
1k<$<23,999				
1981		(not calculated)		
1982	13.62	10.20	0.06	Percentages are
1983	13.41	10.13	0.05	decreasing

DIVORCED

INCOME	MALE	FEMALE	MISSING	TRENDS?
-1<$<4k				
1981		(not calculated)		
1982	1.46	1.80	<0.01	Percentages are
1983	1.47	1.88	<0.01	increasing
0<$<3,999				
1981		(not calculated)		
1982	0.12	0.53	<0.01	Percentages are
1983	0.16	0.59	<0.01	increasing

MARRIED

INCOME	MALE	FEMALE	MISSING	TRENDS?
-1<$<19k				
1981		(not calculated)		
1982	20.44	4.30	0.01	Percentages are
1983	20.59	4.88	0.01	increasing
0<$<3,999				
1981	1.19	9.96	0.04	
1982	1.15	9.60	0.04	Percentages are
1983	1.15	9.42	0.03	decreasing
4k<$<7,999				
1981	2.48	4.93	0.02	
1982		(not calculated)		Percentages are
1983	2.30	4.65	0.02	decreasing
8k<$<18,999				
1981		(not calculated)		
1982		(not calculated)		
1983	9.21	9.11	0.01	

SEPARATED

INCOME	MALE	FEMALE	MISSING	TRENDS?
-1<$<19k				
1981		(not calculated)		
1982	1.46	1.39	<0.01	Fairly stable
1983	1.42	1.41	<0.01	
0<$<3,999				
1981		(not calculated)		
1982	0.14	0.56	<0.01	Fairly stable
1983	0.14	0.58	<0.01	

NOTE: The income ranges for the widow persons were all different for the 3 years and thus no table is given.

Table C.6

GROSS AND NET INCOMES

Nature of income	GROSS=NET=0	GROSS>NET>0	GROSS>0,NET>0
Business			
1981	94.62	3.46	1.35
1982	94.71	3.37	1.37
1983	94.71	3.42	1.33
Professional			
1981	98.78	0.97	0.13
1982	98.74	0.99	0.15
1983	98.73	1.01	0.13
Commission			
1981	99.38	0.29	0.06
1982	99.35	0.31	0.06
1983	99.38	0.29	0.05
Farming			
1981	97.07	1.69	1.01
1982	97.05	1.69	1.04
1983	97.14	1.64	1.01
Fishing			
1981	98.68	0.22	0.04
1982	99.66	0.25	0.04
1983	99.68	0.24	0.04
Rental			
1981	94.49	2.42	2.36
1982	94.35	2.57	2.41
1983	94.13	2.90	2.26
TRENDS	"Rental" shows the largest ↓ over time. "Business" shows the largest ↑ over time.	"Rental" shows largest ↑. Others are fairly stable.	"Rental" is the most variable category.

NOTE: Other types of gross and net income classifications are not dealt with because of the very small counts. Hence, the column totals representing the nature of income will not sum to 100%.